Challenges in Building Usable Knowledge in Education

Larry V. Hedges
Northwestern University

Accepted for Publication in Journal of Research on Educational Effectiveness (2017)

The research reported here was supported by the Institute of Education Sciences, U.S. Department of Education, through Grant **R305D140019** to Northwestern University. The opinions expressed are those of the authors and do not represent views of the Institute or the U.S. Department of Education.

ABSTRACT

The scientific rigor of education research has improved dramatically since the year 2000. Much of the credit for this improvement is deserved by Institute of Education Sciences (IES) policies that helped create a demand for rigorous research; increased human capital capacity to carry out such work; provided funding for the work itself; and collected, evaluated, and made available the results of that work through the What Works Clearinghouse. Major challenges still remain for education research, however. One challenge is dealing with the replication crisis that has plagued other scientific fields and is likely to be a problem in education science. A second challenge is better supporting the generalizability of education research. A third challenge is adapting our rigorous research designs to the increasing complexity of our interventions and our questions about the mechanisms by which these interventions achieve their effects. Promising approaches to meet each of these challenges are suggested.

Challenges in Building Usable Knowledge in Education

In this paper, I provide reflect on the recent history of educational research (where we have come), the current status of education research (where we are now), and most importantly, the challenges that face us as a field whose objective is to build usable knowledge in education (where we have to go). To talk about where we have come from, it is essential to identify a starting point in the recent past. I pick the year 2000, the millennium, as a somewhat, but not entirely, arbitrary starting point. It was the year before the No Child Left Behind Act was passed (in 2001) and two years before the Education Sciences Reform Act that created the US Institute of Education Sciences (IES) was signed into law (in 2002).

Where We Came From

This was a time when education research had, as the distinguished historian of education (and my former colleague) Carl Kaestle wrote, an "awful reputation" (1993). Karl was a sympathetic observer of education research. He was more reporting the opinion of others that he largely did not share, but there was reason for the low esteem in which education research was held. In some parts of education research, the prevailing methodological standards were very low in comparison to other social sciences. There was even a legitimate intellectual position that held that no generalizable knowledge was possible in education (see, e.g., Cronbach, 1975, Cronbach, 1982).

There were also areas where a great deal of progress was being made. Cognitive science that was revolutionizing our understanding of reading (see, e.g., Miller, 2003 or Olson, 2007) and research on education of special populations was helping to create better live for children with special needs (see Blanton, Pugach, and Boveda, 2014). Education research methodology was making important advances. Foundational work on multilevel statistical models (Bryk and Raudenbush, 1992), meta-analysis (Hedges and Olkin, 1985), and psychometrics (e.g., Hambleton, Swaminathan, and Rogers, 1991) was not just thriving but having a huge impact on other fields. No doubt part of the problem was the very amorphous nature of education research (what this term includes and excludes is never clear). Having lived through this era as a professor of education, I recall it as a low point for education research. Internally, we were in the midst of the paradigm wars and the position that no generalizable knowledge was possible in education was respectable, and maybe modal among education school faculty, usually dressed up with some impressive sounding post positivistic rhetoric, (see, e.g., Gage, 1989). Quantitative research training programs and even whole schools and departments of education were closing or under threat, even at the most prestigious universities (e.g., the University of Chicago).

But, at the same time, globalization was underway and education itself was being recognized as important for national economic competitiveness (see, e.g., Heckman, et al., 1999). There was increasing interest in education research from actors outside of the conventional education research realm (from economics, sociology, and public policy studies). Researchers in education schools seemed ill prepared to compete with the challenges brought by the colonization of their territory by other social scientists (particularly economists). Sometimes they even seemed unaware of the implications of having no coherent narrative about what is known and even what is ***worth*** knowing about education. This characterization is impressionistic, but I know that many aspects of it were shared by others at the time.

The No Child Left Behind act was passed in 2001 and the Education Sciences Reform Act was passed in 2002, which created the Institute of Education Sciences (IES). The remarkable character of Grover (Russ) Whitehurst also entered the national scene as the first director of IES. These ingredients brought a sea change in educational research. The NCLB act

jump started a demand for rigorous education research by mandating that rigorous research findings were necessary to justify certain expenditures on education interventions, products, and services. The Education Sciences Reform act created IES, and IES created the What Works Clearinghouse (WWC) to provide an arbiter of which education interventions, products, and services had rigorous research backing. Thus they created a demand for rigorous research and created the first of what became several dissemination mechanisms. Moreover they created funding streams dedicated to *producing* education research through the National Center for Education Research and the National Center for Special Education Research. This was not just a matter of announcing grant competitions and shoveling money out the door. Essential to this effort was reforming the process of reviewing research grant proposals, and establishing standards for IES products and processes for assuring compliance with those standards. The formation and maintenance of a strong standards and review office is one of the unheralded ingredients to IES's success. To increase the capacity of the field to carry out rigorous research in education, IES funded predoctoral and postdoctoral training with the explicit aim of bringing researchers from other fields into education research. They also increased capacity by funding research training for established professionals though summer institutes on randomized field trials, quasi-experimentation, and single case research. Finally, IES provided limited financial support a fledgling professional society called Society for Research on Educational Effectiveness (SREE) whose goals were consistent with those of IES. The idea for SREE originated outside of IES (all of the founders were academic scholars), but I think it is fair to say that all three of the founders (myself included) were inspired by what IES was trying to accomplish.

While one might recount the history of the formation and impact of IES in purely institutional terms, I think it would be a mistake to omit the profound influence that Russ Whitehurst had during his 6 years as IES Director. He had an unswerving vision, extraordinary political skill, enormous energy, and a passion to create an education science that could take its place alongside the disciplines like psychology. He also had a thick enough skin to endure withering criticism from the education research establishment that vilified him. He had an eye for talent that led him to build a competent staff dedicated to his vision of creating a rigorous education science. While I am usually skeptical of "great man" explanations of history, I doubt that as much could have been accomplished as fast at IES without Whitehurst or someone else with his unusual combination of attributes.

Where We Are Now

Looking back at the condition of education research in 2000, I believe that we have come a long way. Then virtually no school of education taught regular courses on randomized trials or rigorous quasi-experimental designs. Today, it is hard to be a serious research oriented school of education without them. Then, few established education researchers had training in doing large scale randomized experiments or high quality quasi-experiments (such as regression discontinuity designs), today literally hundreds have received training through IES sponsored summer institutes. Then few researchers outside of research firms had experience doing randomized trials or high quality quasi-experiments, today hundreds do. The What Works Clearinghouse has lived up to its name, no longer the What ***Doesn't*** Work Clearinghouse. After more than a decade, SREE is alive and well as a professional society to support the work of those who are engaged in rigorous education research. The SREE journal, the *Journal of Research on Educational Effectiveness* is prospering (its first impact factor placed number 3 in impact among over 200 education journals).

A more subtle indicator of progress can be seen in the attitudes of the professional bureaucracy of the federal government and its emphasis on rigorous evaluation. I emphasize these civil servants because, while political appointees come with much fanfare (and usually go with much less), the civil servants are the backbone of federal administration and a very stabilizing force. For example, officials from the Office of Management and Budget have spoken in presentation at the SREE meetings about promoting more rigorous evaluation in government and IES's help in that work, and other OMB officials have argued in other contexts that education (and IES in particular) was considerably ahead of other government agencies and that IES was helping others improve. Other agencies are emulating some of the strategies of IES, including the formation of four WWC-like research clearinghouses in the Departments of Labor, Justice, and Health and Human Services. This is a complement to the professional staff of the department (and especially those in IES), but it is also an indication that rigor has attained a degree of institutionalization at IES and elsewhere that will make it resistant (but not immune) to change, even if future political appointees may wish it to.

I should add that I have emphasized randomized trials as an *indicator* of the rigor of research in our field, not the *definition* of rigorous research (see, e.g., Shavelson and Townes, 2002). Most education research studies are, and will continue to be, things other than randomized trials. This is as it should be. But trials do occupy a special place in the pantheon of study designs that attempt to estimate causal effects and their frequency is an indicator of rigor at one end of the research process.

In 2000, I never imagined that the state of education research would be what it is today. If someone had told me in the year 2000 that they could see the future and had described the state of education research in 2016, I would have been ready to declare we had reached nirvana and my work was done. As a field, we should collectively appreciate how far we have come.

Where We Are Going

After that appreciation, we must move rapidly to a more candid assessment of where we stand today and what the future may hold for us. Our work is nowhere near done. If you have any doubt of that, two recent reports should give us pause. One was the recent Government Accountability Office report on IES, which gave it high marks for rigor but insisted that it needed to work to ensure that its work is more relevant (Government Accountability Office, 2013). Another report was published by the American Enterprise Institute (Smith and Ginsburg, 2016). It raises similar questions of relevance of the evidence in the WWC, and by implication all of IES. The relevance question is important, but in a broader context I see three major challenges that education research must start to address in the next decade and a half. None are easy and but all must be addressed if we are to build usable knowledge from education research.

The most daunting challenge we face is the problem of making our science transparent and replicable. The scientific replicability crisis engulfed medicine, then psychology, and it ***will*** threaten education research. It is not a matter of ***if*** but ***when*** and how will we respond. This challenge is not just a potential annoyance but an existential challenge to science as we know it (and out cushy lives as researchers).

The second general challenge is to better identify the scope of applicability of research findings. We must treat generalizability as a serious validity problem that deserves the same attention that we give internal validity.

Finally, our interventions are becoming increasingly complex, multifaceted, and conditional (think of response to intervention, and tiered intervention programs, for example). The third problem I see involves adapting our development and evaluation methodology to the

complexity of the interventions we need to test. We must give much greater attention to the problem of construct validity of cause in complex interventions, to identify which component bundles are effective and which sequences of treatments are most effective. I will now address each of these challenges in detail.

The Replication Crisis

Biomedical science encountered a crisis of the Kuhnian proportions beginning in the early 2000's. It was a crisis of increasing evidence that biomedical research studies, including large scale randomized trials, were not replicating as expected. We make special claims about the status of scientific knowledge and replicability underlies those knowledge claims (McNutt, 2014). A crisis of replication in science is no less than an existential threat to science itself. This is a threat our special status in society and our claims on government's (which of course means, the citizens') money to support our research.

John Ioannidis was one of the first to call serious attention to the problem in medicine. In 2005, John published two important papers, one modestly titled "Why most published research findings are false" the other with the snappy title "Contradicted and initially stronger effects in highly cited clinical research." The second of these looked at 49 highly cited papers (papers cited more than 1,000 times) in major medical journals from 1990 – 2003 and looked for subsequent replications. He found that 32% of the findings were either contradicted by subsequent studies or found larger effects than subsequent studies did. Only 44% were clearly replicated.

This study itself was replicated and extended in various ways by John and others. This line of work (now called meta-research) has burgeoned. It has become a special interest of *PLOS Biology* and there is a major research center at Stanford Medical School (the METRICS Center) dedicated to it. The problem appears to exist not just in human clinical trials but also in preclinical studies and in animal studies. For example, Silberberg, Schlange, and Asadullah (2011) reported that nearly two-thirds of published preclinical studies that they tried to replicated could be replicated. Similarly Perrin (2014) reported that attempts by the Amyotropic Lateral Sclerosis Therapy Development Institute to replicate published animal studies were stunningly unsuccessful. Figure 1, reproduced from their paper shows that replication attempts for 5 of 9 compounds failed to even replicate the direction of the effect. This line of work had a profound effect at the National Institutes of Health (see Collins and Tabak, 2014), the medical research community, and made it into the popular literature like *Time* magazine and *The Economist* (see the October 19, 2013 issue). Serious research on the replicability of research has edged closer to education. John himself has begun studying social sciences (which so far seems to mean psychiatry) which he claims looks worse than mainstream clinical medicine. John gave keynote address in 2014 at the annual meeting of SREE which we would be wise to understand as a call to action.

Insert Figure 1 About Here

Others have been particularly concerned about replicability in psychology. For example, the Center for Open Science is pursuing a line of work to attempt to replicate major findings in psychology and the results have not been encouraging (see Open Science Collaborative, 2016). Research on replication and replicability in psychology has been increasing, even a special issue of *Psychological Science* was devoted to the replication crisis psychology (see, e.g., Pashler and Harris, 2012).

There has also been a similar concern about reproducibility of results in experimental economics. The Experimental Economics Reproducibility Project attempted to reproduce 18

experimental results published between 2011 and 2014 in the *American Economic Review* and the *Quarterly Journal of Economics* (Camerer, et al., 2016). The results of this reproducibility project also raised concerns that experiments were not as reproducible as expected.

In evaluating the evidence about replication, particularly about replication in the social sciences, it is important to recognize that there is no generally accepted *definition* of replication, something acknowledged by both the Open Science Collaborative (2016) and Camerer (2016). This means that precise analyses and decision procedures with known decision theoretic properties have not been applied to the problem. I believe the right approach for much of social science is one based on meta-analytic ideas: Replication should be defined in terms of parameters representing effects obtained in studies. Accepting this does not eliminate ambiguity. Do studies replicate only if effect parameters (not sample estimates) are *exactly* the same? This criterion is too stringent even in the physical sciences, where some variation in effects is taken to be tolerable (see Hedges, 1987; Olive, et al., 2014; Rosenfeld, 1975). If not, what amount of variation in results should be accepted as being consistent with replication?

Clearly there are moderators of treatment effects that may vary across studies related to samples, contexts, and research procedures. We may try to control them, document them and make them transparent to increase the likelihood of replication. But it seems unlikely that we will be able to fully specify or control them all. Even in the physical sciences, close analysis of attempts to replicate studies show that this can be very difficult because of tacit knowledge that is sometimes required to obtain the same results. For example, in 1970, one laboratory reported that it had constructed a transversely excited atmospheric pressure CO_2 (TEA) laser. The sociologist of science H. M. Collins studied the attempts of 11 laboratories to replicate their results (to build a TEA laser). He found that the help of the scientists who conducted the original study was necessary to replicate their work: "no scientist succeeded in building a laser using only information found in published or other written sources" (Collins, 1985, p. 55). Not only that, "no scientist succeeded in building a TEA-laser where their informant was a 'middle man' who has not built a device himself" (p. 55). Moreover, "even where the informant had built a successful device, . . ., the learner would be unlikely to succeed without some extended period of contact with the informant" (p. 55). I give this example not to argue that there is no replicability crisis, but merely to remind us that it may have many causes.

There are some obvious possible causes of the replicability crisis that *can* be addressed. One is simple chance associated with sampling on the dependent variable. If we select studies by the size of the *estimated* effects they observe (such as Ioannidis' highly cited studies), we are bound to overestimate the true size of the effect. You get a big estimate of effect due to a combination of big true effect plus a lucky (meaning positive) estimation error. That lucky estimation error is unlikely to be repeated in a replication so a smaller estimate is likely to be observed. A related cause could be publication bias in which studies that produce statistically significant effects are more likely to be published and thus published studies may overestimate the true effect size which would be estimated in a replication (see, e.g., Hedges, 1984). There could also be differences in procedures, context, or population between the original study and replication attempts that affect the outcome of studies so that the replications attempts are not actually strict replications of the original study. Failures to replicate would therefore say more about robustness of the original result than about its internal validity.

There are other far more pernicious possible causes that are aspects of research method. I will follow the convention and refer to these as *questionable research practices*. They include changing the hypotheses after the data have been analyzed. This could mean identifying the

primary dependent variable only after analyzing the data on several dependent variables. It could mean identifying hypotheses about which subgroups to look at only after analyzing the data on several subgroups. It could mean identifying which covariates to use after a specification search. It could mean choosing the followup period to use as the definition of the endpoint after having knowledge about the results at various endpoints. It could even mean choosing a transformation, an outlier deletion strategy, or an imputation strategy after knowing what their impact on the findings would be. And there are even more dubious possibilities.

Note that none of these practices involve actually falsifying data. But they all can have profound impacts on the results of the data analysis and they can render statistical hypothesis testing (as well as standard errors and confidence intervals) next to meaningless. In fact these practices have come to be called "*p*-hacking" in psychology (Simmons, Nelson and Simonsohn, 2014). They are clearly not entirely unknown in other fields as well.

The reader might be asking: What well trained and responsible scientist would do things like this? The evidence about that is a little disconcerting. A 2012 study by John, Loewenstein, and Prelec surveyed over 2,000 psychologists about their involvement in ten questionable research practices. They were asked both about whether they admitted using these practices themselves and to estimate the percentage of their colleagues who engaged in these practices. Figure 2, reproduced from their paper provides the principal results. The figure reports the percentage of respondents admitting to each research practice, the estimated prevalence of the practices by respondent's colleagues and a synthetic prevalence estimate. Not surprisingly the respondents appear to admit engaging in these questionable research practices less than they say their colleagues do. A cynic might argue that the true prevalence of these practices is somewhere between what people say *they* do and what they say *their colleagues* do, but even the rates of these practices that the respondents claim to describe their own behavior are quite disturbing. Do anywhere near 40% of psychologists really decide to exclude data after looking at the impact of doing so? I hope not and I hope that education scientists do not.

Insert Figure 2 Here

It is worth noting that this survey found higher rates of admission to questionable research practices than some others, and the survey on which it was based has been criticized because the respondents may not have properly interpreted some of the questions. For example, lower rates of questionable research practices were found by Fiedler and Schwartz (2015) in a survey of German psychologists that used more refined versions of the questions about questionable research practices that were arguably less ambiguous than the questions asked by John, Lowenstein, and Prelec. The results are summarized in Figure 3 (reproduced from Fiedler and Schwartz (2015). The precise text of questions used to elicit information in surveys is obviously important. However one could quibble about "Have you ever done this?" ***versus*** "How often have you done this?" is a better question. Obviously these are *different* questions. Although I admire some of these critics, I worry that scientists would interpret these practices as referring to anything other than questionable scientific practices. Even the German data of Fiedler and Schwarz is not reassuring. If I believe that over 40% of scientists admit having engaged in selective reporting of results that worked, and 25% of them have done it more than once, I am not reassured to hear that they only do it in 10% of their papers. As Fiedler and Schwarz put it "Yes, some researchers admit to faking and lying, though they claim to have done so for only a small subset of the hypotheses they tested"(p. 50). These findings are about psychologists, but you have to wonder how different educational researchers might be.

Insert Figure 3 Here

What I find most disturbing about this study is what it suggests about norms of research methods. These practices (with the exception of falsifying data) could seem harmless to the uninitiated. Yet they completely undermine hypothesis testing as a discipline to control our own enthusiasm for our research hypotheses and therefore the scientific method attached to it. If we are training young scientists that these are acceptable methods, our field is in trouble. But public perception is arguably as important as reality. If the public begins to believe that we are not policing ourselves and that science is not self-correcting (as our rhetoric claims) our field is in trouble.

The problem of course is that these practices are invisible, but need to be discouraged. Clearly we need to address the culture of research methods in education science. It is hard, but not impossible to change culture. The last 20 years provide evidence that some change is possible. Medicine has faced this problem and some of the other social sciences are now beginning to do so. The National Institutes of Health are acting vigorously (see Collins and Tabak, 2014), the National Science Foundation is beginning to (see Bollen, Cacioppo, Kaplan, Krosnick,, and Olds, 2015), and IES needs to formulate its own response.

In medicine, an important strategy for protecting against questionable research practices is the requirement that studies intending to draw causal inferences be registered in advance of data collection. Registration involves publishing a description of the trial and a description of the methods and procedures to be used, including data collection and data analysis procedures. Such descriptions are often called a *protocol* for the trial. The idea of registration and publication of protocols is to create transparency. It require scientists to publicly declare, in advance of data collection, what they intend to do. Ideally, protocols and registration of studies is accompanied by reporting standards that ensure transparent reporting and make it possible to check if important aspects of the plan reported in the protocol were actually implemented in the research being reported.

Note that registration of protocols should not be restricted just to randomized trials. Quasi-experiments can and should have registered protocols too. For example, the *Journal of Observational Studies* recognizes this point and welcomes protocols for observational studies. In fact, Ioannidis (2005) found that nonrandomized studies were less likely to be replicated than randomized trials. This is not surprising because quasi-experiments generally require more complex analytic procedures involving more points of judgment. Thus one can argue that registered protocols may be more important for studies that do not use randomization than for randomized trials.

Most readers probably know about the CONSORT movement that has developed reporting standards, checklists (see Moher, 1998 or http://www.consort-statement.org/), and explanations and elaborations documents (Moher, et al., 2010) that illustrate how to use the standards and checklists. Perhaps the most widely used part of the CONSORT framework is the flow diagram that is frequently used in reporting trials (see Figure 4 is a CONSORT Flow Diagram for a study I will describe later). CONSORT and its variants (especially the variant for cluster randomized trials and social policy interventions) should play an increasing part of education science. There has been a parallel movement in medicine to codify the material that should be reported in protocols for trials: the SPIRIT movement. SPIRIT stands for **S**tandard **P**rotocol **I**tems: **R**ecommendations for **I**ntervention **T**rials. Like Consort, SPIRIT has a checklist (see Chan, et al., 2013b or http://www.spirit-statement.org/publications-downloads/) and explanations and elaborations documents (Chan, et al., 2013a) to facilitate use.

Insert Figure 4 about here

A skeptic could be asking whether these measures have helped in medicine. The jury is still out on that, but there is a small amount of relevant evidence. If questionable research practices produce falsely statistically significant results, then reducing the prevalence of those practices through registration of protocols might be expected to reduce the frequency of trials producing statistically significant results. Bob Kaplan and Veronica Irvin published a paper comparing the proportion of significant results of National Heart Lung and Blood Institute trials in the years before and just after, 2005 when registration became mandatory (see Kaplan and Irvin, 2015). It is interesting that the frequency of significant results among named primary outcomes declined precipitously, while the proportion of significant results among all outcomes was unchanged. This is at least consistent with the hypothesis that registration reduced the prevalence of questionable research practices in medicine.

SREE has taken the lead in addressing the replicability problem in education. With the support of a grant from IES, SREE is developing a registry for education related studies that intend to draw causal conclusions. By the time that this article is published, the SREE registry should be operational. It will permit the user to register a protocol for the study, to update that protocol while retaining the original material, and eventually to incorporate results.

You might be asking, culture change is hard, how will they promote the use of the registry. Obviously education of researchers is a key component in our strategy. In medicine, the journals were a major help. The problem of encouraging registration of trials is bigger than any one journal or even any one professional society. International Committee of Medical Journal Editors (ICMJE) met and agreed to require registration of trials as a prerequisite of publication in their journals. In 2005, an editorial appeared in ICMJE journals, like the New England Journal of Medicine, that announced the policy in bold terms: ”The ICMJE member journals will require, as a condition of consideration for publication, registration in a public trials registry.” Registration became a requirement for publication of clinical trials in high prestige journals. Major funders also helped encourage registration. Similar strategies will help in education too.

It is probably a mistake to think that protocols and registration alone will eliminate questionable research practices and solve the replication problem. The norms of research practice will have to be changed. One model for this might be the effort by the National Institutes of Health to support responsible research conduct and to require that researchers who receive NIH funding have training in responsible conduct of research (see Steneck, 2007).

The Challenge of Understanding Generalizability

Now I want to turn to a different challenge facing education science. A decade ago, I served with several other educational researchers on an American Educational Research Association (AERA) task force to create standards for reporting education research, which were eventually relabeled standards for reporting education research based on social science methods (American Educational Research Association, 2006). This was an interesting experience, in that the task force included both qualitative and quantitative researchers, but we were eventually able to agree on what I thought was a pretty good document. One element of the reporting standards was that a report of research should articulate the scope of applicability of the research findings and a logic by which the findings should apply to that scope. We did not opine about what the scope should be or what the logic connecting the research to the scope of applicability should be. I still think that these considerations are reasonable expectations of any research report.

A few years ago Robinson, et al. (2013) proposed what seemed like an eminently sensible proposition. It seems that several education research journals require every paper they publish to include education policy recommendations (I guess to makes them more "policy–relevant"). The paper about which I was asked to comment argued that not every piece of education research was directly policy relevant and shouldn't have to pretend that it was. Talk about undisciplined thinking—what logic would extrapolate a tiny study designed to answer a development question to broad education policy? Is there any logic here at all? This may be an extreme example, but I think the logic about when a research finding ought to apply to a different setting is pretty primitive.

Let me offer a contrasting example from research in medical education. You may be aware that the training of medical and surgical residents and interns is an arduous process involving long hours of training and patient care. In 1981, then 1989, in 2003, and again in 2011, the Accreditation Council for General Medical Education (ACGME) imposed duty hour limitations on residents and interns in various clinical fields, including surgery. However there has been debate over whether these duty hour restrictions were too strenuous. The arguments were about continuity of patient care, effectiveness of training, and humanness to trainees. In fact, the arguments have gone on for more than 50 years in medical education, but there had been no randomized trials. I helped design the **F**lexibility **I**n duty hour **R**equirements for **S**urgical **T**rainees Trial (known to its intimates as the FIRST Trial) to answer questions about whether offering some flexibility in these requirements would compromise patient care or intern/resident education (see Bilimoria, et al., 2016). The treatment was allowing residency programs to permit some flexibility in implementing duty hour requirements as illustrated in this slide. The details of trial are unimportant to the present discussion, but it is an illuminating contrast to the situation Robinson wrote about. Just to terrorize graduate students and post docs, you might be interested in the duty hour requirements for surgical interns and residents given in Figure 5—it may look a lot like your life now.

Insert Figure 5 About Here

There are 252 ACGME accredited general surgery residency programs. 116 of them were excluded from our sampling frame for various reasons, leaving 136 residency programs. Our sample was 118 residency programs and their 154 affiliated hospitals. That is about 50% of the total residency programs and 87% of the eligible programs. Remember the ACGME makes policy for the 252 residency programs. We did a trial to inform policy that had almost 50% of those programs. Figure 4 shows the CONSORT flow diagram for the study and Figure 6 shows the results of the trial (flexible duty hours was not worse, either in terms of patient outcomes or resident satisfaction).

Insert Figure 6 About Here

I introduce the FIRST trial not because of it outcome but because of the reaction to it by ASGME. I thought this trial had a pretty good sample for addressing the policy question about setting duty hour requirements. However, the first question—the very first question—asked by ASGME when we briefed them on the results was, "Do these results apply to the programs that were not in the study?" My first thought was, "We got half of the population in the sample, what do you want?" My second thought was, "I am glad we anticipated this by planning a study of generalizability to the programs that were not part of the trial" (Chung, et al., in press). My point here is that this is the question sophisticated policy makers ***should*** be asking us, every time we present results to them and they ought to demand some serious scientific warrant for our claims that "everything will generalize just fine."

Of course probability (random) sampling is the gold standard method for generalizing to a specified inference population. Probability sampling is rare, but not entirely non-existent in experimental work. But even where it has been carried out, it has proven very difficult, and I would argue that it is not a panacea for improving generalization from trials. Let me be clear: When it is possible to do probability sampling without compromising either the design of the trial or the definition of the inference population, I advocate probability sampling. For example, in certain technology interventions, random sampling may be relatively easy to implement and it should be used. However there are reasons to think that probability sampling will not be the most frequent technique used to support generalizability in experimental field trials. First it imposes another layer of difficulty and time involved in obtaining cooperation of sites and a necessity of obtaining cooperation from ***every*** site sampled. I suspect that people who have tried to do this would agree that this is hard enough in congressionally mandated studies, let alone investigator initiated studies.

It is also clear that studies doing probability sampling have had (for very good reasons) to impose restrictions on the population definitions used as sampling frames. For example, two of the experiments I know that used probability sampling (and that had a congressional mandate) restricted the sample to programs that were oversubscribed so that random assignment would not deprived of service *because of the trial*. One was the National Head Start Study (Puma, et a., 2010) and the other was the National Evaluation of the Upward Bound Program (Myers and Shirm, 1999). The decision to define the inference populations as they did is a perfectly sensible *practical* decision but it changes the population definition from all programs to oversubscribed programs, which might not be the inference population of most policy interest.

In some cases, trials are used to inform multiple policy questions. In such cases, dual or multiple frame samples can be constructed, but these are more complex than single frame samples and their complexity adds to the complexity of the concomitant experimental designs.

There are some generalization problems that probability sampling cannot solve, even in principle. For example, sometimes there are not just multiple possible inference populations, but they are not known in advance when the experiment is being designed. This happens, for example, when an experiment inspires potential policy changes that had not been anticipated prior to the experiment. (This happened in reaction to the results of the Tennessee Class Size Experiment—other states took note of the results and considered their own class size reductions, see Nye, Hedges, and Konstantopoulos, 2000.)

In the last decade, there has been considerable progress in the logic for applying research findings to specific contexts. This work has been spearheaded by a group of young researchers including Beth Tipton, Liz Stuart, Rob Olsen, and Wendy Chan as well as a few older researchers like Colm O'Muircheartaigh and me (see, e.g., Olsen, et al., 2013; O'Muircheartaigh and Hedges, 2014; Stuart, et al., 2011, Tipton, 2013; Tipton, Hedges, Hallberg, and Chan, 2016). These scholars have given professional development workshops at the meetings of professional societies (e.g., AERA and SREE). A variety of methods have been proposed but all have several features in common.

One obvious point is that if there is no heterogeneity of treatment effects, there is no generalization problem: If treatment effects are always the same, then any old sample will do to estimate the treatment effect. We need to know a lot more about heterogeneity of treatment effects and the variables that predict that heterogeneity. Fortunately, this problem has attracted the attention of a lot of top researchers. In the meantime it seems unwise to bet that treatment effect heterogeneity does not exist.

When heterogeneity does exist, it makes no sense to talk about whether an experiment is generalizable or not without specifying a specific inference population. "Are these results generalizable?" is not a well formed question until you say, generalizable *to what population*. Experimental results may generalize quite well to one inference population, but very poorly to others.

We should be talking about estimates of treatment effects. The experiment that does not have a random sample from the inference population will estimate the average treatment effect in a population of *some* composition. If the inference population has a different composition than the study sample, the average treatment effect in the inference population is likely different from that in the study sample. Thus a better formulation of the generalization question is something like "how well can this experiment support estimation of the average treatment effect in a population with the same composition of the inference population?" The answer can range from "The average treatment effect estimate has a standard error only a little larger than that in the study population (it generalizes pretty well), to "The standard error of the treatment effect estimate in the inference population is infinite" (it offers no insight at all about the treatment effect in the inference population without further assumptions). The later can easily happen when the study sample just doesn't include important parts of the inference population, that is when there is population undercoverage.

The methods currently available and those that are likely to be developed involve modeling selection. They require relevant data about the inference population to allow us to match the study sample to the inference population. The methods, unlike probability sampling, are *not* model free. This means that we need to have to knowledge about heterogeneity and its correlates to develop valid models to support the estimation of treatment effects in policy relevant inference populations.

In pitching this work, I am not arguing that the procedures available are perfect. What I am saying is that they offer coherent logics for supporting a proposed scope of application. Hand waving is an alternative, but many of the hand waving arguments I have heard sound a lot like the intuitions that are formalized by the work of the authors I mentioned previously. Taking seriously the problem of generalizability is an important response to the relevance criticism facing IES and the entire field of rigorous education research.

For the record, the surgical residency programs in the FIRST trial sample *were* systematically different from the surgery training programs that did not participate, but there was very considerable overlap among those populations, leading to credible estimates that flexible duty hours would not result in inferior outcomes in those programs. In other words, we could develop estimates of the average treatment effects in the inference population that had standard errors only modestly larger than those in the study sample, and the estimated effects were not very different than those in the study sample (Chung, et al., in press).

Matching our Research to the Complexity of Our Interventions

Relevance involves testing interventions that are important today (not 10 years ago), but also testing them in ways that give insight into mechanism. We must give much greater attention to the problem of construct validity of cause in complex interventions, to identify which component bundles are effective and which sequences of treatments are most effective. This kind of work fits naturally into the development goals rather than efficacy and effectiveness goals, but sometimes there is ambiguity. The findings about treatments may still need to be tested with a conventional randomized trial, but my point is that more complex interventions and trial designs should have a more important role than they do now. New thinking about designs

that are better suited to developing and evaluating complex treatments would be most welcome and I expect that designs will evolve in conjunction with intervention. Some examples are already available and deserve to be used more than they are. I give two as example, but this is not an exhaustive list of available designs.

One strategy that is particularly useful for developing complex treatments with many components is the **M**ultiphase **O**ptimization **ST**rategy or MOST trial (Collins, Murphy, Nuir, and Strecher, 2005). In fact MOST trials often use standard experimental designs (albeit designs not usually used in education field trials), they are just ideas that are not often used in education science outside the laboratory. Proponents of the MOST strategy would describe a three phase process. A preparation phase where ideas for treatment components are identified and treatment components are designed. Then there is an optimization phase where the components are tried out in various combinations. Typically a fractional factorial design is used to reduce the number of combinations of components that need to be tested. Note that a full factorial design is really not needed here because the object is to find the optimal combination of treatment components, not the completely unambiguous effect of every individual component. The Opt-IN trial for weight loss conducted by my colleague Christine Pelligrini and colleagues (Pellegrini, Hoffman, and Collins, 2014) illustrates the idea. One goal of this trial is to find the optimal treatment subject to a cost constraint of $500/patient. There were five treatment components and the object was to see which combinations were most effective. There are $2^5 = 32$ possible treatment combinations. They used a one-half fractional factorial design that used 16 of those combinations. Such trials do not necessarily need enormous sample sizes to be informative. This trial has not been completed yet but I think you could imagine treatment components like these that could be used in education trials.

Many of our intervention strategies are essentially adaptive interventions that involve multiple possible treatment components that are applied conditionally. It is often useful to use trial designs specifically intended to evaluate adaptive interventions. Adaptive interventions are often created for dealing with great heterogeneity: treatments that might work for some individuals but not for others or what works now may not work as well later, where lack of adherence or high burden is common, or where treatment intensity is required to vary. Adaptive interventions involve some sequencing of decisions regarding treatment, a set of decision points at which treatment options are available, a set of tailoring rules used to trigger changes in treatments, and a sequence of decision rules that use information from the tailoring variables, to make decisions about treatment assignment at each decision point. Adaptive interventions are designed to give insight into critical issues like what sequencing of treatments is best, what timing of alternation of treatments is best.

Sequential Multiple Assignment Randomized Trials (SMART) are multi-stage clinical trials; each participant proceeds through stages of treatment (Collins, Murphy, and Strecher, 2007). Each stage begins with a critical decision and a randomization to treatment takes place at each critical decision. The goal of trial is to inform the construction of an adaptive intervention. The use of the word adaptive in connection with SMART trials might be confusing those who are familiar with more conventional adaptive trials in which the *design* is adaptive (e.g., by changing randomization probabilities as a function of intermediate outcomes). The word adaptive in conjunction with SMART trials refers to the fact that the *intervention,* not the trials design, is adaptive.

An example is a trial conducted by Bill Pelham and Greg Fabian on school interventions for ADHD (Pelhan, et al., 2016). Two basic approaches exist: contingency management and

stimulant medication, used separately or in combination. Opinions vary on the relative merits of each alternative. There isn't much research on the questions of proper dosing and sequencing for a particular child. But those in the psychiatric field tend to favor a medication-based approach, while most parents preferred to start with behavioral treatments.

In this trial, children were randomly assigned at the beginning of the school year to receive low-dose medication or to receive low-intensity behavior modification with parental involvement. After 8 weeks, the children whose treatment was working continued with that treatment, subject to monthly reassessment and re-randomization if deterioration occurred. All others were randomized a second time. For those originally in the behavior modification group ("BehFirst"), some had medication added ("B-then-M" group); others had behavioral modification intensified ("B-then-B" group). For those originally in the medication group ("MedFirst"), some had behavior modification added ("M-then-B" group); others had medication intensified ("M-then-M" group). Figure 7 summarizes the design. The study was directed at seeing how student behavior in the classroom could be improved. They discovered that behavioral interventions followed by intensified behavioral interventions had a much better response than behavioral interventions followed by medication. The worst result—by a large margin—occurred when initial medical intervention was followed by behavioral intervention. This is practically useful information.

Insert Figure 7 About Here

Many variants of these two ideas are possible (see Ktsanes, 2017) and some variants need not use random assignment at every point. Use of these ideas can lead to the development of better and more relevant interventions. It can also lead to experimental evaluations that provide more insight about the components of treatments and their interactions with each other. Education science needs MOST trials, SMART trials, the variety of conventional randomized trials, and strong quasi-experiments to build a foundation of usable knowledge in education.

Note that matching our research designs to the complexity of our interventions may also involve approaches that involve longer term collaborations with practitioners aimed at sequential improvement of educational systems of the kind that the Carnegie Foundation for the Advancement of Teaching has advocating (Bryk, Gomez, and Grunow, 2011). The essential challenge is how to use improvement science ideas while providing rigorous evaluations of the improvements so that continuous change is actually continuous improvement.

Conclusions

Education science has accomplished a great deal in the last 15 years by increasing the number of studies with rigorous designs, especially but not exclusively, randomized trials. This suggests that the internal validity of our evaluation studies has improved. There is reason to believe that the education scientists are currently better trained than in the past and training programs are poised to continue to produce new researchers whose training will improve the human capital stock of education science. Yet education science faces very significant challenges in the future

While evidence has been widely respected in the federal government and elsewhere for the last 15 years, it need not always be so. Evidence is often inconvenient, particularly when it challenges values or passionately held beliefs. The education science community must join hands with the entire scientific community to support the importance of scientific evidence both for its own sake and to inform wise public policy. It will be important to do so not by just standing firm in defense of science and our business as usual. We need to educate and to persuade, but we also need to learn ourselves. Not every skeptic is entirely uninformed, nor are

they all enemies of reason. We need to learn about why others may be skeptical of our evidence and avoid the temptation to be condescending. We also need to remember that evidence is important in informing policy decisions, but it is not the only component of wise policy decisions—extra-scientific considerations such as values and preferences also must play a role.

In our interactions with those outside the scientific community, humility has an important role to play. Being clear about what we have evidence about and what we do not, and how certain that evidence might be, will serve us well. Overstatement of our findings or our certainty about them, is dangerous to our credibility (and it should be). This is part of why a decisive response to the replication crisis is important. Inability to replicate findings and widespread questionable research practices undermine the credibility of not only our scientific claims but also our claims to the societal resources that support our field.

More attention to the generalizability of our research findings should inspire us to temper our claims about the applicability of the evidence we do have. The difficulty is to recognize that while we can make *some* strong claims based on evidence, but the broader the reach of the generalization, the weaker those claims are likely to become. We need to be able to articulate this while making the case that what we know is useful, even if it is limited.

Honesty and clear communication about what we know will only takes us so far. We need to improve the reach of our evidence, pushing our science in directions that are more relevant to practical problems of education. It is not an accident that the improvement science approach advocated by the Carnegie Foundation for the Advancement of Teaching has been so attractive to local education agencies. It promises a disciplined approach to the practical problems facing education today and it is not incompatible with rigorous testing of improvements via innovative research designs. A constructive engagement with that movement could infuse more rigor in their evaluations and more relevance in the rigorous research I hope we will pursue.

Finally, we should be aware of the challenges and opportunities posed by the big data revolution. I am skeptical that administrative and other incidental data collections can entirely replace designed data collections (surveys and experiments), but more data, particularly universe data can be helpful in a myriad of ways we probably don't understand yet. It clearly helps with understanding and improving generalizability of studies, but also can help with identification of gaps in our knowledge and in generating hypotheses.

References

American Educational Research Association (2006). Standards for Reporting on Empirical Social Science Research in AERA Publications. *Educational Researcher, 35*, 33-40.

Bilimoria, K. Y., J. W. Chung, L. V. Hedges, A. R. Dahlke, R. Love, M. E. Cohen, D. B. Hoyt, A. D. Yang, J. L. Tarpley, J. D. Mellinger, D. M. Mahvi, R. R. Kelz, C. Y. Ko, D. D. Odell, J. J. Stulberg, and F. R. Lewis. (2016). "National Cluster-Randomized Trial of Duty-Hour Flexibility in Surgical Training." *The New England Journal of Medicine, 374*, 713-27.

Blanton, L. P., Pugach, M. C., & Boveda, M. (2014). *Teacher education reform initiatives and special education: Convergence, divergence, and missed opportunities (Document No. LS-3*). Retrieved from University of Florida, Collaboration for Effective Educator, Development, Accountability, and Reform Center website: http://ceedar.education.ufl.edu/tools/literature-syntheses/

Bollen, K., Cacioppo, J. T., Kaplan, R. M., Krosnick, J. A., and Olds, J. L. (2015). *Reproducibility, replicability, and generalization in the social, behavioral, and economic sciences.* Report of the Subcommittee on Replicability in Science Advisory Committee to the National Science Foundation Directorate for Social, Behavioral, and Economic Sciences. Arlington, VA: National Science Foundation.

Bryk, A. S., Gomez, L. M., & Grunow, A. (2011). Getting ideas into action: Building networked improvement communities in education. In M Hallinan (Ed.) *Frontiers in Sociology of Education*. New York: Springer Publishing.

Bryk, A. S. & Raudenbush, S. W. (1992). *Hierarchical linear models*. Newbury Park, CA: Sage Publications.

Camerer, C. F., Dreber, A., Forsell, E., Ho, T., Huber, J., Johannesson, M., Kirchler, M., Almenberg, J., Altmejd, A., Chan, T., Heikensten, E., Holzmeister, F., Imai, T., Isaksson, S., Nave, G., Pfeiffer, T., Razen, M., Wu, H. (2016). Evaluating replicability of laboratory experiments in economics. *Science, 351*, 1433-1436.

Chan, A-W., Tetzlaff, J. M., Gøtzsche, P. C., Altman, D. G., Mann, H., Berlin, J., Dickersin, K., Hróbjartsson, A., Schulz, K. F., Parulekar, W. R., Krleža-Jerić, K., Laupacis, A., Moher, D. (2013a) SPIRIT 2013 Explanation and Elaboration: Guidance for protocols of clinical trials. *British Medical Journal*, *346*,e7586.

Chan A-W., Tetzlaff, J. M., Altman, D. G., Laupacis, A., Gøtzsche, P. C., Krleža-Jerić, K., Hróbjartsson, A., Mann, H., Dickersin, K., Berlin, J., Doré, C., Parulekar, W., Summerskill, W., Groves, T., Schulz, K., Sox, H., Rockhold, F.W., Rennie, D., Moher, D. (2013b) SPIRIT 2013 Statement: Defining standard protocol items for clinical trials. *Annals of Internal Medicine,158*, 200-207.

Chung, J. W., Bilimoria, K. Y., Stulberg, J. J., Quinn, C. M., & Hedges, L. V., (in press). The Estimation of Population Average Treatment Effects in the FIRST Trial: Application of a Propensity Score-Based Stratification Approach, *Health Services Research.*

Collins, H. M. (1985). *Changing order.* London: Sage.

Collins, F. S. & Tabak, L. A. (2014). NIH plans to enhance reproducibility, *Nature, 505*, 612-613.

Collins, L., Murphy, S. A., & Strecher, V. (2007) The multiphase optimization strategy (MOST) and sequential multiple assignment randomized trial (SMART): New methods for more potent ehealth interventions. *American Journal of Preventive Medicine, 32(5 Supplement)*, S112-S118.

Cronbach, L. J. (1975). Beyond the two disciplines of scientific psychology. *American Psychologist, 30*, 116-127.

Cronbach, L. J. (1982). *Designing evaluations of educational and social programs*. San Francisco: Jossey-Bass.

Gage, N. L. (1989). The paradigm wars and their aftermath: A "historical" sketch of research on teaching since 1989. *Educational Researcher, 18*, 4-10.

Government Accountability Office. (2013). *Education research: Further Improvements Needed to Ensure Relevance and Assess Dissemination Efforts*. Washington, DC: US Government Accountability Office. https://www.gao.gov/assets/660/659425.pdf

Fiedler, K. & Schwarz, N. (2016). Questionable research practices revisited. *Social Psychological and Personality Science*, *7*, 42-52.

Hambleton, R. K., Swaminathan, H., & Rogers, H. J. (1991). *Fundamental of item response theory*. Newbury Park, CA: Sage Publications.

Heckman, J., Cawley, J., Lochner, L., & Vytlacil, E. (1999). Understanding the role of cognitive ability in accounting for the recent rise in the economic return to education. In K. Arrow, S. Bowles, and S. Durlauf (Eds.) *Meritocracy and economic inequality*. Princeton: Princeton University Press.

Hedges, L. V. (1984). Estimation of effect size under nonrandom sampling: The effects of censoring studies yielding statistically insignificant mean differences. *Journal of Educational Statistics, 9*, 61-85.

Hedges, L. V. (1987). How hard is hard science, how soft is soft science?: The empirical cumulativeness of research. *American Psychologist, 42*, 443-455.

Hedges, L. V. & Olkin, I. (1985). *Statistical methods for meta-analysis*. New York: Academic Press.

Ioannidis, J. P. A (2005). Contradicted and iniatilly stronger effects in highly cited clinical research. *Journal of the American Medical Association, 294(2)*, 218-228.

Ioannidis, J. P. A (2005). Why most published research findings are false. *PLOS, Medicine, 2(8)*, e124.

John, L. K., Loewenstein, G., & Prelec, D. (2012). Measuring the prevalence of questionable research practices with incentives for truth telling, *Psychological Science, 23*, 524-532.

Kaestle, C.F. (1993). The awful reputation of educational research. *Educational Researcher, 22 (1),* 23-26-31.

Kaplan, R. M. & Irvin, V. (2015). Likelihood of Null effects of large NHLBI clinical trials has increased over time. *PLOS One*, *10(8)*, e0132382

Katsanes, R. (2017). *Design and analysis of trials for developing adaptive interventions in education*. Evanston, IL: Northwestern University Institute for Policy Research Working Paper.

McNutt, M. (2014). Reproducibility, *Science, 343*, 229.

Miller, G. A. (2003). The cognitive revolution: A historical perspective. *Trends in Cognitive Science, 7*, 141-144.

Moher, D. (1998). CONSORT: an evolving tool to help improve the quality of reports of randomized controlled trials. Consolidated Standards of Reporting Trials. *Journal of the American Medical Association, 279*,1489-1491.

Moher, D., Hopewell, S., Schulz, K. F., Montori, V., Gøtzsche, P. C., Devereaux, P. J., Elbourne, D., Egger, M., & Altman, D. G. (2010). CONSORT 2010 explanation and

elaboration: Updated guidelines for reporting parallel group randomized trials. *British Medical Journal, 340*, c869.

Myers, D.. & Schirm, A. (1999). *The impacts of Upward Bound: Final report for phase I of the national evaluation.* Washington, DC: Mathematica Policy Research.

Nye, B., Hedges, L. V., & Konstantopoulos, S. (2000). The effects of small classes on achievement: The results of the Tennessee class size experiment. *American Educational Research Journal, 37*, 123-151.

Olive, K,A. et al. (2014).Review of particle properties. *Chinese Physics Journal C,38*,090001. http://iopscience.iop.org/issue/1674-1137/38/9

Olson, D. (2007). *Jerome Bruner: The cognitive revolution in educational theory*. New York: Bloomsbury Academic Publishers.

Olsen, R. B., Orr, L. L., Bell, S. H., & Stuart, E. A. (2013). External validity in policy evaluations that choose sites purposively. *Journal of Policy Analysis and Management, 32*, 107-121.

O'Muircheartaigh, C. & Hedges, L. V. (2014). Generalizing from experiments with non-representative samples. *Journal of the Royal Statistical Society, Series C, 63.* 195-210.

Open Science Collaborative (2016). Estimating the reproducibility of psychological science. *Science, 349*, 943-951.

Pashler, H. & Harris, C. R. (2012). Is the replicability crisis overblown? Three arguments examined. *Psychological Science, 7*, 531-536,

Pelham, W. E., Fabiano, G. A., Waxmonsky, J.G., (2016) Treatment sequencing for childhood ADHD: A multiple-randomization study of adaptive medication and behavior interventions. *Journal of Clinical and Adolescent Psychology*, *45*, 396-415.

Pellegrini, C. A., Hoffman, S. A., Collins, L., & Spring, B. (2014). Corrigendum to "Optimization of remotely delivered intensive lifestyle treatment for obesity using Multiphase Optimization Strategy: Opt-IN study protocol". *Contemporary Clinical Trials, 38,* 251-259.

Perrin (2014). Make mouse studies work. *Nature, 507*, 423-425.

Prinz, F., Schlange, T., & Asadullan, K. (2011). Believe it or not: How much can we rely on published data on potential drug targets? *Nature Reviews Drug Discovery, 10*, 712-713.

Puma, M., Bell, S. Cook, R., & Heid, C. (2010). Head Start Impact Study. Final Report. Washington, DC: U.S. Department of Health and Human Services, Administration for Children and Families.

Robinson, D. H., Levin, J. R., Schraw, G., Patal, E. A., & Hunt, E. B. (2013). On going (way) beyond one's data: A proposal to restrict recommendations for practice in primary educational research journals. *Educational Psychology Review, 25(2),24-28.*

Rosenfeld, A. (1975) The particle data group: Growth and operations. *Annual Review of Nuclear Science*,

Shavelson, R.J., and Towne, L., (Eds). (2002). *Scientific research in education.* Committee on Scientific Principles for Education Research. Center for Education. Division of Behavioral and Social Sciences and Education, National Research Council. Washington, DC: National Academy Press.

Simonsohn, U., Nelson, L. D., & Simmons, J. P. (2014). The *p*-curve: A key to the file-drawer. *Journal of Experimental Psychology: General*, *143* (2), 534.

Smith, M. S. & Ginsburg, H. (2016). *Do randomized trials meet the "gold standard": A survey of the usefulness of the randomized trials in the What Works Clearinghouse*. Washington, DC: American Enterprise Institute.

Steneck, N. H. (2007). *ORI Introduction to the Responsible Conduct of Research*. Washington, DC: US Government Printing Office.

Stuart, E.A., Cole, S.R., Bradshaw, C.P., and Leaf, P.J. (2011). The use of propensity scores to assess the generalizability of results from randomized trials. *The Journal of the Royal Statistical Society, Series A 174(2),* 369-386

Tipton, E. (2013). Improving generalizations from experiments using propensity score subclassification: Assumptions, properties, and contexts. *Journal of Educational and Behavioral Statistics, 38,* 239-266.

Tipton, E. C., Hedges, L. V., Hallberg, K., & Chang, W. (2016). Implications of small samples for generalization: Adjustments and rules of thumb. *Evaluation Review*, *40*, 1-34.